Harmonizing Instruments with Equating

Singh, Ranjit K.

different organizations, but I predict that WHO, JHU, and Worldometer will emerge as the core cross-national data sources that academics and governments will use. WHO is the sole official source of cross-national data. Yet, JHU quickly established itself as a premier data provider early in the process and has held on to that status. Worldometer has been well-known to social scientists, and thus may be used because their data can easily be merged with the social, economic, and political data that they already provide.

I do not assess the validity and reliability of these data, but I do mention some of the sources of error that may lead to discrepancies across nations and time. A main source is the frequent redefinitions or cases and mortality, which are due in part to changes in knowledge about Covid 19. As countries update their knowledge, it is not clear whether they will expend the effort to retrospectively change their data to reflect the new knowledge. Other errors occur in any large-scale data collection process, such as processing errors.

A source of error that has yet to gain much attention is the difficulty in harmonizing and aggregating data from multiple local sources that report upstream to national organizations. These discrepancies may be due to unequal infrastructures and the unequal resources of hospitals, labs, and other organizations staffed with time and social pressured people who, due to systemic problems and simple fatigue, make mistakes that can introduce a series of minor errors in the data that they report upstream. The upstream reporting problem may not matter much, or it may matter a lot. We don't know. Upstream reporting is a black box that we should open.

Joshua K. Dubrow is co-editor of Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences. *This material is based upon work supported by the National Science Foundation under Grant No. (PTE Federal award 1738502) and by the National Science Centre, Poland (2016/23/B/HS6/03916).*

Harmonizing Instruments with Equating

by Ranjit K. Singh

This is a brief introduction to equating, which is a promising approach to harmonising survey instruments that measure latent constructs such as attitudes, values, intentions, or other individuals' attributes that are not directly observable. The focus is on measurement instruments with only one question (in contrast to multi-item questionnaires). The article is intended to inspire ex-post harmonization practitioners who struggle to harmonize such variables into a homogenous target variable. Some researchers might find that the equating approach is directly applicable to their work. However, even if the specific approach is not a good fit for a particular project, the basic idea of equating could still be a helpful way of thinking about instrument harmonization in general.

The Goal of Instrument Harmonization

A common challenge in ex-post harmonization is how to combine data on a concept measured with different instruments. This is especially hard if the concept is a latent construct; i.e., a construct that cannot be directly observed, such as attitudes, values, or intentions. The central problem here is that latent constructs have no natural units. There is, for example, no self-evident way to compare how much "strongly interested" is on one scale of political interest as compared to "somewhat agree" on another scale measuring political interest.

To get a better understanding of the problem, it helps to clarify the ideal result of an ex-post harmonization process: Taking data from different surveys that were not intended to be combined and that use different instruments, we want to create a seamless dataset. Seamless here implies that once the dataset has been harmonized, it should no longer matter which instrument was used for a particular case in the dataset. To that end, we need to harmonize scores measured with different instruments so that the same score always "means" the same, regardless of the source instrument.

While seemingly self-evident, it is necessary to take a closer look what "meaning the same" implies. To understand harmonization, we should first remind ourselves that the output of measurement (the observed scores on an instrument) is not reality itself, but only something that is related to this reality (Raykov & Marcoulides 2011). What this entails is best explained with the following example.

Consider a latent construct such as political interest. If we measure political interest, we implicitly assume that respondents have a certain attribute strength that governs the extent to which they direct their attention towards or away from sources of political information. In other words, they have a theoretical *true score* that directly reflects their real interest. When we ask respondents to answer a standardized question about their political interest, we assume that respondents will choose one of the offered response categories based on their true political interest. In other words, a measurement instrument projects the true score of respondents onto an arbitrary numerical scale: People within a certain range of political interest will likely chose a "1" on the instrument, people in a somewhat higher range of political interest will likely chose a "2," and so on for each score up to "5" on the five-point scale. The crux is that different instruments project the same reality — true scores — differently. The "3" on one scale does not automatically represent the same range of true scores as the "3" on another instrument. This is true even if the two instruments have the same number of response options but differ with respect to other features, such as wording or layout.

The relationship between true scores and measurement scores in different instruments has two important implications: (1) The *observed scores* we have in our source data represent a mix of truth and measurement; and (2) Different instruments change the measurement component and are likely to result in different observed scores for people with the same true scores (Raykov & Marcoulides, 2011). With this in mind, we can now formalize what instrument harmonization should do: Respondents who are the same with regard to a construct should get the same harmonized score, no matter which instrument was used.

Linking and Equating

Fortunately, this goal of harmonization is shared by psychometric performance and aptitude testing, where there is a need to make comparable different tests for the same construct. This has resulted in an extensive literature dating back to the 1970s on what is today called score linking (Dorans & Puhan, 2017). Equating, meanwhile, is a subfield of score linking that directly addresses our problem, i.e., making comparable scores from instruments that measure the same construct.[1] As we will see, both the logic of equating and its formulas can be of great use.

At the heart of equating is its equity property. Respondents with a certain true score should, on average, get the same converted (i.e., equated) score in a source instrument than they would get on the target instrument (Kolen & Brennan, 2014). The equity property contains the qualification "on average" to reflect random error in measurement and in equating itself. If we achieve such a matching of converted source scores to target scores, we have corrected differences in measurement without eliminating or biasing real differences.

The obstacle we now face is, of course, that we do not have the true scores — we only have observed scores. Psychometry tackles that problem with multi-item instruments such as personality questionnaires. True score estimations are extracted from the interplay of different measurements of the same construct for each respondent, often in factor models (Raykov & Marcoulides, 2011). In the social sciences, large scale survey programs often cannot accommodate multiple questions for all their constructs of interest. Fortunately, not all forms of equating rely on multiple items. Observed score equating relies only on the observed scores (what we have in the dataset) and not on true score estimations. Without multiple measurements, that is, multiple data-points for each person, we cannot disentangle measurement and reality on the respondent level. We can, however, disentangle measurement and reality on the aggregate level.

Observed Score Equating

The basic idea of observed score equating is that if we cannot isolate the effect of different instruments, we control for population differences in true scores via random group designs. This is done by taking two random samples of the same population. In one, respondents answer the source instrument, and in the other, there is the target instrument. Since both samples have similar true scores distributions (they randomly sample the same population), differences in the observed score distributions are due to the instrument differences. Next, we apply a mathematical transformation to

[1] A short note on terminology: In the formal literature, equating also denotes a very strict quality standard for test comparability in psychometric diagnostics, such as professional aptitude testing. This formal standard is only attainable with test forms constructed to very similar test specifications such as length or reliability (Kolen & Brennan, 2014). For harmonization purposes, equating can still be done even if the instruments for example differ in reliability. The standard only cautions us that equating makes units of measurement comparable but does not correct the limitations of the equated instruments. For an extensive overview of the terminology and its history, see Kolen and Brennan (2014) or Dorans and Puhan (2017).

scores of the source instrument so that the distribution of transformed source scores is similar in shape to the distribution of target scores (Kolen & Brennan, 2014). In other words, observed score equating basically matches scores in the two instruments based on their position along the frequency distribution. Average respondents get the same scores regardless of instrument used, and the same is true for below average or above average respondents.

Equating does not mean that the survey data we want to harmonize has to be drawn from the same population. We can use different datasets to perform the equating that results in a transformation table. This table can then be used to harmonize the data we intend to harmonize. Equating is also always symmetrical, meaning that we can transform scores from one instrument to the other and *vice versa*.

Next, we take a closer look at two ways to transform distribution shapes in observed score equating: (1) Linear equating for approximately normally distributed source and target scores and (2) equipercentile equating if the distribution of one or both instrument scores are non-normal (e.g., strongly skewed or even bimodal).

Linear equating

Linear equating assumes that both score distributions are approximately normally distributed, which implies that the two instrument score distributions only differ in two parameters: The mean and the standard deviation. In linear equating, scores of the source instrument are linearly transformed so that the transformed source score mean and standard deviation become equal to the target score mean and standard deviation (Kolen & Brennan, 2014). Respondents now have very similar scores on the transformed source instrument and the target instrument depending on their position along the normal distribution. Respondents with the same z-score have the same harmonized score but scaled to the format of the target scale.

To avoid confusion, I add two clarifications. First, linear equating is distinct from a mere z-transformation. The mathematical transformation that is used to align the distribution shapes is indeed similar to a z-transformation. However, at the heart of linear equating are the two instrument samples drawn from the same population. By setting the population as equal, we can isolate and eliminate the measurement differences. The resulting translation table can then be used in instances where the two instruments were used on non-equal populations. The result is a harmonization of the measurement while preserving true population differences. A mere z-transformation, in contrast, would indiscriminately destroy true population differences because for each sample we will have mean = 0, and SD = 1.

Second, linear equating is also quite different from the frequently used harmonization approach, which is the linear stretch method. Linear stretch applies a linear transformation to instrument scores solely based on differences in scale points. Consider a five-point source scale and a seven-point target scale. With linear stretch, we would assign minimum score to minimum score (a source score of 1 would remain a 1) and maximum score to maximum score (a source score of 5

would become a 7). All points in between are stretched so that they fit in the space between 1 and the new maximum score with equal distances (Jonge, Veenhoven, & Kalmijn, 2017). The source scale 1, 2, 3, 4, 5 would become the transformed scale 1, 2.5, 4.0, 5.5, 7.

Linear equating, in contrast, harmonizes scales based on the distribution of responses for the same population. Linear stretch only takes the number of response scale points into account. The difference becomes apparent if we apply both methods to two instruments with different question and response option wording, but with the same number of scale points. Even in the same population we would expect different response frequency distribution for both instruments because both question wording and response option wording change the response options that respondents choose. Linear stretch would ignore that and assign each score in one instrument to exactly the same score in the other instrument because the scale points are the same (Jonge et al., 2017). Linear equating, meanwhile, would transform scores so that the distributions become aligned. Hence, respondents at the same position along the construct distribution (e.g., average respondents) may well get different harmonized scores with linear stretch, but with very similar scores with linear equating.

Equipercentile Equating

Equipercentile equating, meanwhile, drops the assumption of normally distributed instrument scores. Instead, we transform scores so that the distribution of transformed source scores fluidly matches the shape of the distribution of target scores. And as a reminder: Just like with linear equating, this is performed on two random samples from the same population. Equipercentile equating operates like this: We take a score from the source instrument and based on the frequency distribution we calculate the percentile rank of that score (i.e., the position of that score along the distribution of the construct in the population). Then we look up which score in the target instrument corresponds to that percentile rank. After that, we transform the source score with a certain percentile rank into that target score with the same percentile rank (Kolen & Brennan, 2014). Consequently, all scores now "mean the same" in the sense that each transformed source score and target score point to the same specific place in the population distribution.

One remaining challenge is that response scores are ordinal and not continuous. This has two implications that the equipercentile equating formulas solve with linear interpolation (Kolen & Brennan, 2014). (1) A score does not represent an exact percentile rank along the continuous distribution of the construct. Instead, each score represents a range of respondents (e.g., if the first response option is chosen by 20% of respondents, then it represents percentile ranks from the 0^{th} to the 20^{th}). Equipercentile equating solves this with linear interpolation and simply assigns the middle (e.g., 10%). (2) If we have a percentile rank for a source score, we likely have no target score at exactly that percentile rank. Again, we linearly interpolate and assign a transformed score between the two target scores. If the two applicable target scores 1 and 2 represent the 10^{th} and 40^{th} percentile rank, and if the percentile rank of the source score is 20, then we would assign the

transformed score 1.25. This is because 20% is 33% along the distance from 10% to 40% and 1.33 is 33% along the distance from score 1 to score 2. The outcome of equipercentile equating is then, just as with linear equating, a transformation table that translates scores of one instrument into the format of the other instrument.

Observed Score Equating in Practice

At this point I was hopefully able to pique your interest in observed score equating. Yet the hurdle of requiring two samples from the same population, one for each instrument, poses a challenge. In the following, I will lay out some arguments and ideas how observed score equating can be applied in practice.

First, to stress again, the data used to equate two instruments does not have to be from the dataset we want to harmonize. Any point in time where the two instruments are used in a probabilistic sample of the same population is enough. This might also mean that the matching is done with a completely different survey program that just so happens to have copied one of the instruments. Finding such serendipitous scale-population-time matches is facilitated by using the databases of large harmonization projects, such as the Survey Data Recycling project (and previous project: see Tomescu-Dubrow and Slomczynski 2016), which may well have the data you need conveniently searchable, bundled, and cleaned.

Second, if you have probabilistic samples of the same population, but at different points in time, then this is only a problem if the population changed substantially regarding the construct in the meantime. Chances are that another survey program has a time series of the construct. If no change occurred, equating can be done as is, and if a change occurred, one can apply linear equating but correct the transformation for the change over time.

Third, observed score equating is preferably done with samples from a population that is similar to the population that the harmonization project is interested in. Yet, if no such data exist, equating can also be done with a split-half experiment in a non-probabilistic sample (e.g., an online access panel). With a single survey, several instruments can be equated at the same time. Please note that comparability issues can occur if the experiment sample and the target population of your project are very different and if your measurement instrument is not measurement invariant across those populations (i.e., if it is not interpreted the same way across populations). Dorans and Holland (2000) discuss the problem, the potential ways of estimating the extent of the problem, and ways to mitigate it.

Fourth, with regard to harmonizing cross-cultural data, equating often cannot be done directly with existing data. This is often the case if data from national survey programs is to be harmonized, meaning that the same construct is measured with different instruments in different countries. However, equating can perhaps still be used via chained equating (Kolen & Brennan, 2014). Consider harmonizing instruments from two probabilistic national surveys from Country A and Country B. Now assume that the relevant construct has also been included in a cross-national

survey program including Countries A and B. We can then equate the instruments from the national survey A to the cross-cultural survey. And then equate the cross-cultural instrument further to the national survey B. However, this chained equating cannot be more comparable across cultures than is the instrument in the cross-national survey.

In sum, equating is worth considering. It is not a panacea, but if the preconditions are met, it is a rigorous method that will increase the quality of the harmonized dataset considerably. If suitable data are available, then equating is easily done due to many existing specialized programs and packages for statistical applications. A package specialized in observed score equating is *equate* (Albano, 2016). Packages that also include other techniques are *kequate* (Andersson, Branberg, and Wiberg 2013) and *SNSequate* (González, 2014). For an overview and guidance on applying these R packages see González & Wiberg (2017). There are also stand-alone programs for equating, such as *RAGE-RGEQUATE* for observed score equating (Zeng, Kolen, Hanson, Cui, & Chien, 2005) which can be downloaded at Brennan (2020) who also curates a comprehensive list of other programs.

Yet even if a direct application is not possible, the idea of equating is helpful in rethinking harmonization. It also guides us to potential pitfalls in analysing harmonized data where equating has not been performed. If you would like to learn more about what equating can do for your project or what alternative approaches exist, feel free to contact me. At GESIS, I offer consultation on how to make survey instruments comparable in ex-post harmonization (see below).

Ranjit K. Singh is a post-doctoral scholar at GESIS, the Leibniz Institute for the Social Sciences, where he practices and researches the harmonization of substantive instruments in surveys. At GESIS he now also offers consulting on the ex-post harmonization of substantive instruments for researchers who either combine data from different survey programs or who want to change an instrument in their ongoing survey program. Consultation topics include assessing instrument comparability, weighing consequences of comparability issues, and strategies for harmonizing substantive instruments. For more information, contact ranjit.singh@gesis.org or visit: https://www.gesis.org/en/services/data-analysis/data-harmonization/harmonizing-substantive-instruments

References

Albano, A. D. (2016). equate: An R Package for Observed-Score Linking and Equating. *Journal of Statistical Software*, *74*(8). https://doi.org/10.18637/jss.v074.i08

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software*, 55(6). https://doi.org/10.18637/jss.v055.i06

Brennan, R. L. (2020). *Computer Programs*. Retrieved June 24, 2020, from https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs

Dorans, N. J., & Holland, P. W. (2000). Population Invariance and the Equatability of Tests: Basic

Theory and The Linear Case. *Journal of Educational Measurement*, *37*(4), 281–306.

Dorans, N. J., & Puhan, G. (2017). Contributions to Score Linking Theory and Practice. In R. E. Bennett & M. von Davier (Eds.), *Advancing Human Assessment: The Methodological, Psychological and Policy Contributions of ETS* (pp. 79–132). https://doi.org/10.1007/978-3-319-58689-2_4

González, J. (2014). SNSequate : Standard and Nonstandard Statistical Models and Methods for Test Equating. *Journal of Statistical Software*, *59*(7). https://doi.org/10.18637/jss.v059.i07

Jonge, T. de, Veenhoven, R., & Kalmijn, W. (2017). *Diversity in Survey Questions on the Same Topic: Techniques for Improving Comparability*. https://doi.org/10.1007/978-3-319-53261-5_1

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3rd ed.). https://doi.org/10.1007/978-1-4939-0317-7

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York: Routledge.

Tomescu-Dubrow, I., & Slomczynski, K. M. (2016). Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling. *International Journal of Sociology*, *46*(1), 58–72. https://doi.org/10.1080/00207659.2016.1130424

Zeng, L., Kolen, M. J., Hanson, B.A., Cui, Z., & Chien, Y. (2005). *RAGE-RGEQUATE [computer program]*. Iowa City, IA: The University of Iowa.

Inter-Survey Methodological Variability in Institutional Trust from the Survey Data Recycling Framework

By Joonghyun Kwak

A critical problem in survey data harmonization is methodological inter-survey variability. Methodological differences between surveys have been treated as unmeasured errors that might be inherent in survey question properties or might occur during fieldwork and data processing (Slomczynski and Tomescu-Dubrow 2018). The Survey Data Recycling (SDR) framework offers a solution to this problem by creating harmonization and survey quality control variables that measure potential sources of inter-survey variability. This research note examines the extent to which the harmonization and survey quality control variables adjust for inter-survey variability, using trust measures in parliament, legal system, and political parties from the SDR database v1.1 (Slomczynski et al. 2017).